

DOCUMENT RESUME

ED 285 913

TM 870 521

AUTHOR Mitchell, Karen J.; Anderson, Judith A.
TITLE Reliability of Holistic Scoring for the 1985 MCAT Essays.
PUB DATE Aug 86
NOTE 10p.; Paper presented at the Annual Meeting of the American Psychological Association (Washington, DC, August 22-26, 1986).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Analysis of Variance; College Entrance Examinations; *Essay Tests; Higher Education; *Holistic Evaluation; *Interrater Reliability; Medical Schools; Pretesting; Scoring; *Writing Evaluation
IDENTIFIERS *Medical College Admission Test

ABSTRACT

A pilot essay was included in the 1985 Spring and Fall administrations of the Medical College Admission Test. A sample of 320 of the essays written by Fall examinees who had expressed an interest in allopathic medicine was used to calculate interrater reliability estimates. Sixteen of 20 readers who had been trained by White's suggestions for holistic scoring were included in the reliability study. Each reader rated two batches of 20 essays on the first day and again on the second. Analyses of variance were computed for (1) day; (2) batch; (3) reader; (4) subjects nested within batch; (5) interaction of day with batch; (6) interaction of day with reader; (7) interaction of reader with batch; and (8) interaction of subjects within batch with day. The interrater reliability index was estimated at .84, indicating that 71% of the variation in scores was due to level of differences between the essays themselves. The remainder was due to batch, day, day by reader, day by batch, batch by reader, and reader. When extrapolated to the operational setting, the index was .72. Methods of limiting the effects of these factors are proposed. (MGD)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

ED285913

RELIABILITY OF HOLISTIC SCORING FOR THE 1985 MCAT ESSAYS

Karen J. Mitchell and Judith A. Anderson

Association of American Medical Colleges

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

K. J. Mitchell

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☒ This document has been reproduced as
received from the person or organization
originating it.

☐ Minor changes have been made to improve
reproduction quality.

• Points of view or opinions stated in this docu-
ment do not necessarily represent official
OERI position or policy.

Reliability of Holistic Scoring for the 1985 MCAT Essays

Abstract

The Association of American Medical Colleges is conducting research on the incremental validity associated with adding an essay to the Medical College Admission Test (MCAT). The present research examines the reliability of holistic scoring for a sample of essays written during the Fall 1985 MCAT administration. Analysis of variance techniques are used to estimate interrater reliability and to partition score variance into that due to level differences between papers and to context-specific factors such as rater, time of scoring, etc. Recommendations for structuring and managing large-scale scoring sessions are presented based on the results.

Reliability of Holistic Scoring for the 1985 MCAT Essays

Introduction

The Association of American Medical Colleges (AAMC) is conducting research on the incremental validity associated with adding an essay to the Medical College Admission Test (MCAT). The objectives of the research are to develop, implement, and evaluate an MCAT essay testing program. Evaluation activities for the essay pilot project are described in "MCAT Essay Pilot Project Evaluation Plan" (AAMC, 4 March 1985). The analysis plan for initial stages of the pilot project are described in "Analysis Plan for the MCAT Essay Pilot Project" (AAMC, 17 April 1985).

The essay was administered on a pilot basis with the 1985 Spring and Fall MCAT administrations. The essay prompts were designed to provide examinees with the opportunity to demonstrate skill in: 1) developing a central idea, 2) synthesizing concepts and ideas, 3) separating relevant from irrelevant information, 4) developing alternative hypotheses, 5) presenting ideas cohesively and logically, and 6) writing clearly, observing the accepted practices of grammar, syntax, punctuation, and spelling consistent with timed, first draft composition.

Spring 1985 Administration

After the Spring MCAT administration, twenty experienced scorers met for three days to rate a sample of 3,000 papers written by Saturday examinees. The rater training and scoring sessions were structured using White's (1985) suggestions for organizing and managing holistic essay readings. Each essay was rated on a six point holistic scale by two readers. The two assigned ratings were summed to yield a final score rating from two to twelve. A small number of the papers (n=279) were rated by four readers to provide estimates of interrater reliability.

A number of analyses of variance (ANOVAs) were computed using data from the smaller reliability data set. The ANOVAs indicated that 66% of the variation in scores was due to level differences between the essays themselves. The remainder of the score variance was attributable to context-specific factors such as: 1) the particular readers who scored the papers, 2) the tables at which the readers were grouped, 3) the batches to which the papers were assigned, 4) the days on which the papers were read, and/or 5) other unexamined factors. The rank ordering of these variables in terms of their impact on score variation was batch, table, readers nested within table, and the interaction of batch with table. Design limitations prevented the estimation of rating variation due to day of reading. When the .81 interrater reliability estimate was extrapolated to the operational setting where only two ratings were assigned to each essay, the reliability index became .72.

Fall 1985 Administration

After the Fall 1985 MCAT administration, the scorers reassembled to rate a sample of 1,600 essays written by Saturday examinees. The papers were systematically selected to represent the demographic and academic characteristics of the Fall examinee population. Papers for initial and repeat examinees indicating a preference for allopathic medicine were included in the sample.

The rater training and scoring sessions were, again, structured using White's (1985) suggestions for managing holistic scorings. Essays were randomly batched into groups of twenty and batches were each assigned at random to two readers. Readers were seated at three tables; each table was headed by a table leader. Because the Spring data pointed to a table or table leader effect, table leaders were rotated across tables during the scoring session so that each leader headed all three groups. Scores from the two six-point ratings were, again, summed to provide a score scale ranging from two to twelve.

Sixteen batches ($n=320$) from the Fall testing were randomly selected to provide data for calculation of interrater reliability estimates. Data on these sixteen batches are the primary focus of this work.

Objectives

The objectives of this research are to examine the reliability of holistic scoring for the Fall 1985 MCAT essay and to partition score variation into that due to actual level differences between papers and variance due to context-specific factors such as: batch assignment, date of scoring, reader and the interactions of these variables. Suggestions for structuring holistic scoring sessions are presented based on these data.

Methodology and Results

Sixteen of the twenty readers were randomly selected for inclusion in the reliability study and four batches were randomly assigned to each. Two of the four batches assigned to each reader were rated on the first day of scoring and the remainder were scored on the second day.

Analysis of variance techniques and generalizability theory were used to calculate interrater reliability estimates for the MCAT essay. An ANOVA was computed with the following factors: (1) day, (2) batch, (3) reader, (4) subjects nested within batch, (5) the interaction effect of day with batch, (6) the interaction of day with reader, (7) the interaction of reader with batch, and (8) the interaction of subjects within batch with day. A balanced, repeated measures ANOVA design with one nested factor and four duplications was used (Kirk, 1982). Three-way interactions were pooled with the residual term. Mean squares from the four duplications were averaged. Quasi F-ratios were calculated using the residual sum of squares as the error term for all factors.

Table 1
Fall 1985 MCAT Essay Interrater Reliability Results

<u>Source</u>	<u>df</u>	<u>Duplication 1</u>	<u>Duplication 2</u>	<u>Duplication 3</u>	<u>Duplication 4</u>	<u>Average Mean Square</u>	<u>Quasi F-ratio</u>	<u>Range of Group Means</u>
Date	1	.889	1.253	3.781	.681	1.651	4.12	3.79 - 3.82
Batch	3	1.870	2.615	1.161	3.565	2.303	5.34	3.44 - 4.04
Reader	3	.393	.291	.948	1.370	.7505	1.74	3.54 - 4.01
Subjects (Batch)	68	2.344	2.310	2.544	1.724	2.231	5.18	
Batch x Reader	9	.528	1.13	1.281	.349	.822	1.91	
Date x Batch	3	.611	1.624	.892	.375	.876	2.03	
Date x Reader	3	1.204	.170	1.198	1.681	1.063	2.47	
Date x Subjects (Batch)	68	.303	.448	.409	.444	.401	.93	
Residual	129	.422	.346	.506	.449	.431		
Total	287							

Reliability results for the Fall 1985 MCAT essay are presented in Table 1. The range of group means for each factor also appears in the table. The interrater reliability index for the Fall scoring was estimated at .84. This means that 71% of the variation in scores, as measured by the effect of subjects within batch, was due to level differences between the essays themselves. The remainder was due to the effects of the following factors: batch, day, day by reader, day by batch, batch by reader, and reader. The interrater reliability index was .72 when extrapolated to the operational setting.

Conclusions

These data speak to the reliability of the scoring process as it was configured and suggests revision of scoring procedures for future administrations. Data from the 1985 Spring and Fall scoring indicated that both batch assignment and date of reading had the largest context-specific impact on essay scores. Because in both administrations essays were randomly assigned to batches, the average score differences between batches were highly unexpected. After the Spring administration, it was posited that readers may form a general impression of a batch based on the first few papers in a group and then read the remainder of the batch in light of that impression. At both scorings, however, the papers in a batch were ordered differently for each of the four readers. This means that the readers would have had to form the same early impression of a batch based on different initial papers. The plausibility of this argument can be questioned. The salience of the batch effect is puzzling. A small number of papers were grouped in various sized batches at the Fall reading to examine the impact of possible gestalt-type reading. The same papers were read as part of batches of 10, 20, and 30 to see if score assignment varied by size of batch. No differences were observed between scorings under the different batching conditions.

Differences in mean scores by day suggest that calibration exercises should be conducted with frequency; that is, once the actual scoring begins, raters should be stopped periodically and their scoring checked to see if it is on scale and consistent with scoring guidelines. Perhaps it is also wise to periodically insert activities into large scale readings to combat fatigue or boredom. Results suggest that the two scores for an essay should be assigned on different days to lessen the effect of a particular high or low day.

At the Spring scoring, readers were assigned to one of three tables and their assignments to tables remained constant over the three days. Differences between mean table scores indicated that readers operated as a group in some way, either in concert with other readers or at the influence of the table leaders. Because interaction among readers at a table was generally low, it seemed more likely that there were systematic differences in the ways leaders directed the readings at their tables. Spring results suggested rotating table leaders across tables as the scoring proceeded. The rotation of tables leaders was introduced as a control at the Fall reading and no average differences between tables were observed. Because this particular group of readers and table leaders had worked together extensively, the rotation caused no difficulty for leaders in either getting to know readers' scoring patterns or in developing a rapport with table members. Rotation may be less successful with a less familiar group.

Despite the above discussion of experimental effects, it should be noted that the portion of variance attributable to these factors is not large. The range of group means for readers was 3.54 to 4.01; for date of reading, the range was 3.79-3.82, and the range for batch was 3.44 to 4.04. The standard error of measurement for the eleven-point scale was .90, which is 8.2% of the score scale. This compares to a .87 standard error of measurement for the MCAT science tests which is 5.8% of the fifteen-point MCAT scale. The latter calculation is based on an internal consistency measure of reliability.

The essay reliability indices discussed, here, are interrater reliability estimates. They do not address the reliability of the MCAT essay across topics. They also do not speak to the intrarater or test/retest reliability of the exam. Because it is planned that unique essay questions will be presented at each MCAT administration, it is necessary to establish the equivalence of essay topics across different testings. Score equivalence is necessary both to provide admissions committees with a frame of reference for the data and to facilitate comparison of scores in applicant pools with test data from different MCAT administrations. More importantly, score equivalence is essential to a guarantee of fairness to applicants.

Variations in the format and focus of essay topics will be evaluated using data from the September 1986 administration of the MCAT. Data on the equivalence of results across topics will also be available for a small number of MCAT re-examinees. Fall 1986 data will provide estimates of scoring and test reliability across raters, topics and days. Pilot administration of the MCAT essay will continue through 1987 and data will be used to inform the development and administration of essay prompts that are equivalent across testings. Suggestions for structuring and managing large-scale scoring sessions will be extended and refined based on pilot data.

Bibliography

Analysis Plan for the MCAT Essay Pilot Project (1985, April). Available from Division of Educational Measurement and Research, Association of American Medical Colleges, Washington, D.C.

Kirk, Roger E. (1982). Experimental Design: Procedures for the Behavioral Sciences. Belmont, California: Brooks/Cole Publishing Company.

MCAT Essay Pilot Project Evaluation Plan (1985, March). Available from the Division of Educational Measurement and Research, Association of American Medical Colleges, Washington, D.C.

Mitchell, Karen J. and Anderson, Judith (1986). Reliability of Holistic Scoring for the MCAT Essay. Educational and Psychological Measurement, 46, 3.

White, Edward M. (1985). Teaching and Assessing Writing: Recent Advances in Understanding, Evaluating and Improving Student Performance. San Francisco, California: Jossey-Bass Publishers.